Applying GCS Networks to Fuzzy Discretized Microarray Data for Tumour Diagnosis

Fernando Díaz[1], Florentino Fdez-Riverola[2], Daniel Glez-Peña[2], and J.M. Corchado[3]

[1] Dept. Informática, University of Valladolid, Escuela Universitaria de Informática,
Plaza Santa Eulalia, 9-11, 40005, Segovia, Spain
fdiaz@infor.uva.es
[2] Dept. Informática, University of Vigo, Escuela Superior de Ingeniería Informática,
Edificio Politécnico, Campus Universitario As Lagoas s/n, 32004, Ourense, Spain
{riverola, dgpena}@uvigo.es
[3] Dept. Informática y Automática, University of Salamanca,
Plaza de la Merced s/n, 37008, Salamanca, Spain
corchado@usal.es

Abstract. Gene expression profiles belonging to DNA microarrays are composed of thousands of genes at the same time, representing the complex relationships between them. In this context, the ability of designing methods capable of overcoming current limitations is crucial to reduce the generalization error of state-of-the-art algorithms. This paper presents the application of a self-organised *growing cell structures* network in an attempt to cluster biological homogeneous patients. This technique makes use of a previous successful supervised fuzzy pattern algorithm capable of performing DNA microarray data reduction. The proposed model has been tested with microarray data belonging to bone marrow samples from 43 adult patients with cancer plus a group of six cases corresponding to healthy persons. The results of this work demonstrate that classical artificial intelligence techniques can be effectively used for tumour diagnosis working with high-dimensional microarray data.

1 Introduction and Motivation

In recent years, machine learning and data mining fields have found a successful application area in the field of DNA microarray technology. Microarrays are one of the latest high-throughput technologies in experimental molecular biology, which allow monitoring of gene expression for tens of thousands of genes in parallel and are already producing high amounts of valuable data. One of the major uses of DNA microarray experiments is to attempt to infer meaningful relationships among genes, but the analysis and handling of such data is becoming one of the major bottlenecks in the utilization of this technology [1]. Since the number of examined genes in an experiment is measured in terms of thousands, different data mining techniques have been intensively used to analyse and discover knowledge from gene expression data [2]. However, having so many fields relative to so few samples creates a high likelihood of finding false positives.

Recent studies in human cancer have demonstrated that microarrays can be used to develop a new taxonomy of cancer, including major insights into the genesis, progression, prognosis, and response to therapy on the basis of gene expression profiles [3].

E. Corchado et al. (Eds.): IDEAL 2006, LNCS 4224, pp. 1095 – 1102, 2006.

However, there continues to be a need to develop new approaches to (*i*) diagnose cancer early in its clinical course, (*ii*) more effectively treat advanced stage disease, (*iii*) better predict a tumour's response to therapy prior to the actual treatment, and (*iv*) ultimately prevent disease from arising through chemopreventive strategies. Given the fact that systematic classification of types of tumours is crucial to achieve advances in cancer treatment, several research works have been developed in this direction [4, 5].

Following a novel approach, in this paper we explore the capabilities of a growing cell structures (GCS) neural network to discover relevant knowledge for clustering patients suffering for acute myeloid leukemia (AML). This knowledge can be easily conveyed to and understood by humans via available visualization techniques. A key advantage of the proposed method is that it allows incorporating biological meaningful information to the network operation in the form of a gene-based distance metric. Our GCS network makes use of a previous successful fuzzy discretization method for data reduction on microarray data domain.

The rest of the paper is organized as follows: Section 2 summarizes our previous successful fuzzy discretization algorithm for data dimensionality reduction, Section 3 presents the main issues about the proposed model based on a modified GCS network, Section 4 presents the experiments carried out and discusses the obtained results, finally, Section 5 gives out the concluding remarks.

2 Discovering Relevant Genes Using a Discriminant Fuzzy Microarray Descriptor

Input space reduction is often the key phase in the building of an accurate classifier [6]. Based on a novel fuzzy discretization method working as the retrieval stage in the GENECBR system [7], it is possible to represent any microarray by means of its generalized fuzzy microarray descriptor (FMD). This descriptor is a comprehensible representation for each gene expression level in terms of one from the following linguistic labels: LOW, MEDIUM and HIGH. Moreover, from a set of FMDs the method is also able to construct a prototype, known as a fuzzy pattern (FP), which characterizes and summarizes the most relevant values of gene expression levels within a given set of microarrays. A FP is a higher concept constructed from a set of FMDs. A fuzzy pattern can be viewed as a prototype of the set of FMDs from which it is constructed. Therefore, the fuzzy pattern can capture relevant and common information about the gene expression levels of these FMDs. The final goal of the proposed method is to select a reduced number of relevant and representative genes allowing other artificial intelligence techniques being able to tackle with this high-dimensional domain.

As aforementioned, the proposed method employs a fuzzy codification for the gene expression levels of each microarray, based on the discretization of real gene expression data into a small number of fuzzy membership functions. The whole algorithm comprises of three main steps: (*i*) first, we discretize the gene expression levels into binary variables according to a supervised learning process generating several FMDs; then, (*ii*) a unique FP is generated from the patients belonging to each specific pathology; finally, (*iii*) we discriminate between those genes belonging to the existing FPs and we select a subset of relevant genes in order to construct the final discriminant fuzzy pattern (DFP). When the algorithm finishes each microarray can be represented

by a simplified fuzzy vector of common genes that we call FMD_{DFP}. The details of the whole process can be found in [8].

In this work we plan to employ this new representation of the available samples (FMD_{DFP}) as input data for a GCS network. The target goal of the GCS network is to group those patients that are most similar to a new one but only taking into account the genetic information provided by the previously selected genes (DFP vector descriptor).

3 GCS Networks for Clustering Biologically Homogeneous Patients

GCS neural networks [9] constitute an extension to Kohonen's self-organising maps [10], and are only one member in the family of self-organising incremental models. GCS networks have the advantage of being able to automatically construct the network topology, and to support easy visualisation of semantic similarity in high-dimensional data. More importantly, the extracted knowledge that is relevant to clustering can provide meaningful explanations for the clustering process and useful insight into the underlying domain.

To illustrate the working model of the GCS network used in our experimentation, a two-dimensional space is used, where the cells (neurons) are connected and organized into triangles [11]. Each cell in the network is associated with a weight vector, w, of the same dimension as the number of relevant genes selected in the previous step (size of the DFP vector). At the beginning of the learning process, the weight vector of each cell is initialized with random values [11]. The basic learning process in a GCS network consists of topology modification and weight vector adaptations carried out in three steps.

In the first step of each learning cycle, the cell c, with the smallest distance between its weight vector, w_c, and the actual FMD_{DFP} is chosen as the *winner cell* or best-match cell. The selection process is succinctly defined by using the Euclidean distance measure as indicated in Expression (1) where O denotes the set of cells within the structure at a given point in time.

$$c : \left\| FMD_{DFP} - w_c \right\| \leq \left\| FMD_{DFP} - w_i \right\|; \forall i \in O \tag{1}$$

The second step of the learning process consists of the adaptation of the weight vector, w_c, of the winning cell, and the weight vectors, w_n, of its directly connected neighbouring cells, N_c, by means of Equations (2) and (3).

$$w_c(t+1) = w_c(t) + \varepsilon_c (FMD_{DFP} - w_c) \tag{2}$$

$$w_n(t+1) = w_n(t) + \varepsilon_n (FMD_{DFP} - w_n); \forall n \in N_c \tag{3}$$

where ε_c and ε_n represent the learning rates for the winner and its neighbours respectively, belonging to the [0, 1] interval, and N_c stands for the set of direct neighbour cells of the winning cell, c.

In the third step, a *signal counter*, τ, is assigned to each cell, which reflects how often a cell has been chosen as winner. Equations (4) and (5) define how the signal

counter is updated with parameter α acting as a constant rate of counter reduction for the rest of the cells at the current learning cycle, *t*.

$$\tau_c(t+1) = \tau_c(t) + 1 \tag{4}$$

$$\tau_i(t+1) = \tau_i(t) - \alpha\,\tau_i(t); i \neq c \tag{5}$$

Growing cell structures also modify the overall network structure by inserting new cells into those regions that represent large portions of the input data (genetically similar patients), or removing cells that do not contribute to the input data representation. The cell deletion policy has not been used in our work due to the lack of great amounts of data. The adaptation process is then performed after a fixed number of learning cycles of input presentations (epochs). Therefore, the overall structure of a GCS network is modified through the learning process by performing only cell insertion. Equations (6), (7) and (8) define the rules that govern the insertion behaviour of the network.

$$h_i = \tau_i / \sum_j \tau_j ; \forall i, j \in O \tag{6}$$

$$q : h_q \geq h_i ; \forall i \in O \tag{7}$$

$$r : \left\| w_r - w_q \right\| \geq \left\| w_p - w_q \right\| ; \forall p \in N_q \tag{8}$$

Insertion starts with selecting the cell, which served the most often as the winner, on the basis of the signal counter, τ. The cell, *q*, with the highest relative counter value, *h*, is selected. The neighbouring cell, *r*, of *q* with the most dissimilar weight vector is determined using Expression (8). In this expression, N_q denotes the set of neighbouring cells of *q*. A new cell, *s*, is inserted between the cells *q* and *r*, and the initial weight vector, w_s, of this new cell is set to the mean of the two existing weight vectors, w_q and w_r. Finally, the signal counters, τ, in the neighbourhood, N_s, of the newly inserted cell, *s*, are adjusted. The new signal counter values represent an approximation to a hypothetical situation where *s* would have been existing since the beginning of the process.

An important issue in the network operation is the distance calculation between two cells or one cell and the actual FMD_{DFP}. Every time the network needs to compute a distance between two nodes, Expression (1) is used. In this work, we propose to hybridise our GCS network by using biological knowledge for the distance computation. Given that each cell in the network have a weight vector, w_c, representing their location in the input space (FMD_{DFP} space) and that each position in w_c stands for a gene expression value, we can use the similarity between linguistic labels (represented by fuzzy sets) as a measure of the relation between each point belonging to w_c and the corresponding value in the FMD_{DFP} vector.

In order to explain how we calculate this correspondence we need to previously define the similarity between linguistic labels (represented by fuzzy sets). In this case, it has been considered that the fuzzy intersection of two fuzzy sets *A* and *B* (represented by its membership functions, μ_A and μ_B, respectively) is given by the application of the `min` operator to the two membership functions, namely, $\mu_{A \cap B} = \texttt{min}\,\{\mu_A, \mu_B,\}$. On the other hand, the cardinality operator can be replaced by the integral operator (see

[8] for details). In this way, the metric $Sim(A, B)$ varies between the values 0 (total dissimilarity) and 1 (total similarity).

It is important to highlight that the final goal of our GCS network is to cluster all patients that are genetically similar given a selected group of genes (DFP vector descriptor) and without taking into account their previous assigned classes. Our proposed method aims to find new relations between the patients even now unknown. Therefore, it is possible and not contradictory to group together patients suffering different (but genetically related) diseases. Such a topology has the added advantage that inter-cluster distances can be precisely quantified. Since such networks contain explicit distance information, they can be used effectively to (*i*) represent an *indexing structure* which indexes sets of related patients and to (*ii*) serve as a similarity measurement between individual patients.

Every time a new microarray needs to be classified a new FMD_{DFP} is constructed and presented to the trained CGS network. A sorted vector of pairs, S, holding the similarity of each selected patient with the new microarray is generated. In order to produce a new classification, a proportional weighted voting schema is proposed. For this purpose, we need to ponder the vote of each patient contained in vector S. In this case, a weight α_j for each retrieved patient, k_j, is calculated based on the position (pos) that it occupies in the vector S and the level of similarity with the target case, Sim_j. For this task, Expression (9) is used.

$$\alpha_j = Sim_j \frac{2^{|S|-1}}{(2^{|S|}-1)2^{pos-1}} \tag{9}$$

Therefore, the classification made by the GCS network when a target patient is presented to the system depends on both the number of selected patients (those genetically similar taking into account the genes belonging to the DPF vector descriptor) and the level of similarity with the target patient. The solution proposed by the system is the class corresponding to the disease with the highest score.

As we can surmise, it is easy to introduce a rejection mechanism in the voting model. We simply use a threshold T to indicate whether the score received by the best matching class is sufficiently strong (passing quota). In the event that the score received by the matching class is less than T, the target patient remains unclassified.

4 Evaluation

The goal of this section is to evaluate the performance of the GCS network in conjunction with the dimensionality reduction technique based on the notion of FMDs. The GCS is trained and tested over an available set of 49 microarrays from the Haematology Service of the University Hospital of Salamanca (Spain).

Acute myeloid leukemia is a heterogeneous group of hematological cancers with marked differences in their response to chemotherapy. As in many other human cancers, the diagnosis and classification of AML have been based on morphological, cytochemical and immunophenotipic features. More recently, genetic features have helped to define biologically homogeneous entities within AML [12]. Karyotype is the most important independent prognostic factor and therefore the most useful parameter for stratifying patients into risk groups. Thus, the favorable outcome group is

composed of well-defined subtypes in terms of cytogenetics: t(15;17), inv(16) and t(8;21) [13, 14]. In contrast, the correlation between morphologic characteristics, genetic abnormalities and prognostic features is more inconsistent in the remaining AML. Analysis of gene expression profiles of tumors using microarray technology has become a powerful tool for classifying hematopoietic neoplasms [15].

Bone marrow samples from 43 adult patients with newly de novo diagnosed AML were analyzed. All samples contained more than 80% blast cells. The median age was 36 years (range 14-70 years). Patients were classified according to the WHO classification into 4 subgroups: (*i*) 10 APL with t(15;17) confirmed by FISH studies with LSI PML/RARA probe (Vysis, Stuttgart, Germany), (*ii*) 4 AML with inv(16) confirmed by FISH analysis with LSI CBFB probe (Vysis); (*iii*) 7 acute monocytic leukemias and (*iv*) 22 non-monocytic AML without recurrent cytogenetic translocations. In addition to this data, a set of 6 samples from healthy persons are also available and they constitute the group of control. Each case (microarray experiment) stores 22,283 expressed sequence tags (ESTs) corresponding to the expression level of thousands of genes. The data consisted of 1,091,867 scanned intensities.

The employed methodology splits the available data within a test set and a training set with 1/3 and 2/3 of the whole observations, respectively. In order to asses the maximum number of nodes of the GCS network an strategy of cross-validation is used, concretely a three fold stratified cross validation. In each round, each fold of the original training set is used to estimate the predictive accuracy of the GCS network which has been trained from the rest of folds. The mean error (and its standard error) for each number of nodes of the GCS are depicted in Figure 1, which shows the training and evaluation errors. As it can be seen, the configuration of 6 nodes as maximum number of cells of the GCS networks involves the minimal value of the error in the evaluation sets, so this value was used to train a GCS network from the training set.

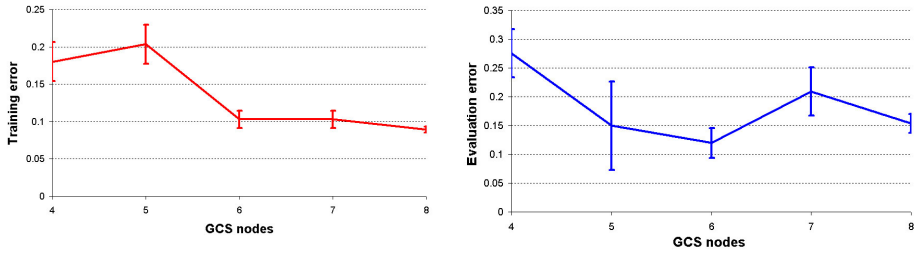

Fig. 1. CV#3 error of GCS network v.s. maximum number of cells at the parameter estimation

Before the training of the GCS network, a reduced number of features (genes) are selected using the FMD representation of the available microarrays. Specifically, the original number of 22,283 ESTs per microarray was reduced to only 165 meaningful ESTs. After the training of the GCS network (with a maximum number of cells equal to 6) over the training set (with 34 observations) the classification error of the GCS network over training was a 8.82 % and a 6.67% over the test set (with 15 observations).

Finally, the following considerations can be made from the clustering performed by the GCS network. The control group (samples from healthy persons) can be adequately differentiated from patients with any kind of AML (see Figure 2). The APL and AML-mono groups are also well differentiated from the rest of AML groups, and there is some kind of overlapping among the AML-inv and the AML-other groups. It must be remembered that the AML-other group is the uncertain area at the current knowledge in the field of AML. The APL group is a kind of AML well characterized (morphologically, cytogenetically and genetically), whereas the AML-mono or AML-inv are possible kinds of AML which are partially characterized. Finally, the AML-other is no characterized in any way. Therefore, at the present state-of-the-art, the given classification can present mistakes and it is possible that some samples from the AML-other group belongs to the AML-mono, AML-inv or new subtypes of AML.

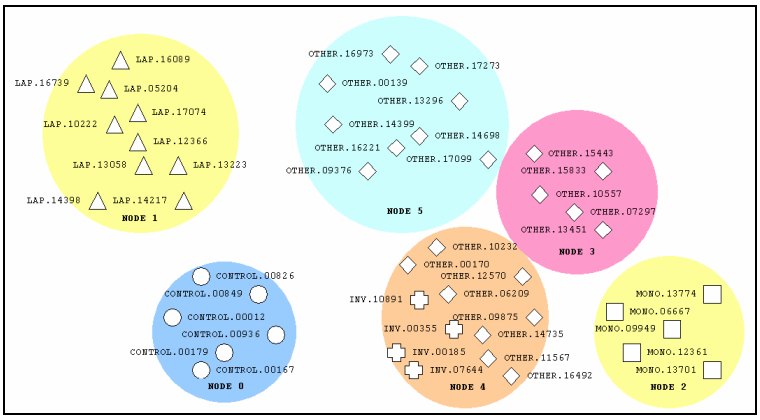

Fig. 2. Final mapping between patients and cells in the GCS network space

5 Conclusions

This work explores the capabilities of a growing cell structure neural network to discover relevant knowledge for clustering patients suffering for acute myeloid leukemia. A key advantage of the proposed method is that it allows incorporating biological meaningful information to the network operation in the form of a gene-based distance metric. Moreover, the GCS network makes use of a previous successful fuzzy discretization method for data reduction on microarray data domain.

Using self-organising GCS networks to meaningfully cluster filtered microarray data has a number of appealing features over other approaches. For example, incremental self-construction, and easy visualisation of biological relationships among the input data. The explanations of the clustering process carried out by the network can be addressed by means of our DFP vector. The most relevant knowledge for each cluster can be highlighted, and provide meaningful explanations about the clustering process and useful insight into the underlying problem and data. The experimental results show that with only a small subset of the genes belonging to a microarray, the performance of the network in terms of the clustering accuracy rate raises to 100%.

References

1. Piatetsky-Shapiro, G., Tamayo, P.: Microarray data mining: facing the challenges. ACM SIGKDD Explorations Newsletter, Vol. 5 (2). (2003) 1–5
2. Cho, S.B., Won, H.H.: Machine learning in DNA microarray analysis for cancer classification. Proc. of the First Asia-Pacific Bioinformatics Conference, (2003) 189–198
3. Ochs, M.F., Godwin A.K.: Microarrays in Cancer: Research and Applications. BioTechniques, Vol. 34. (2003) s4–s15
4. Xiang, Z.Y., Yang, Y., Ma, X., Ding, W.: Microarray expression profiling: Analysis and applications. Current Opinion in Drug Discovery & Development, Vol. 6 (3). (2003) 384–395
5. Golub, T.: Genome-Wide Views of Cancer. The New England Journal of Medicine, Vol. 344. (2001) 601–602
6. Cakmakov, D., Bennani, Y.: Feature selection for pattern recognition. Informa Press, (2003)
7. Díaz, F., Fdez-Riverola, F., Corchado, J. M.: GENE-CBR: a Case-Based Reasoning Tool for Cancer Diagnosis using Microarray Datasets. Computational Intelligence. ISSN 0824-7935. In Press.
8. Fdez-Riverola, F., Díaz, F., Borrajo, M.L., Yáñez, J.C., Corchado, J.M.: Improving Gene Selection in Microarray Data Analysis using fuzzy Patterns inside a CBR System. Proc. of the 6th International Conference on Case-Based Reasoning, (2005) 191–205
9. Fritzke, B.: Growing Self-organising Networks – Why?. Proc. of the European Symposium on Artificial Neural Networks, (1993) 61–72
10. Kohonen, T.: Self-Organising Maps. Springer-Verlag, (1995)
11. Fritzke, B.: Growing Cell Structures - A Self-organizing Network for Unsupervised and Supervised Learning. Technical Report, International Computer Science Institute, Berkeley, (1993)
12. Vardiman, W., Harris, N.L., Brunning, R.D.: The World Health Organization (WHO) classification of the myeloid neoplasms. Blood, Vol. 100. (2002). 2292–2302
13. Grimwade, D., Walker, H., Oliver, F., Wheatley, K., Harrison, C., Harrison, G., Rees, J., Hann, I., Stevens, R., Burnett, A., Goldstone, A.: The importance of diagnostic cytogenetics on outcome in AML: analysis of 1,612 patients entered into the MRC AML 10 trial. Blood, Vol. 92. (1998) 2322–2333
14. Slovak, M.L., Kopecky, K.J., Cassileth, P.A., Harrington, D.H., Theil, K.S., Mohamed, A., Paietta, E., Willman, C.L., Head, D.R., Rowe, J.M., Forman, S.J., Appelbaum, F.R.: Karyotypic analysis predicts outcome of preremission and postremission therapy in adult acute myeloid leukemia: a Southwest Oncology Group/Eastern Cooperative Oncology Group Study. Blood, Vol. 96. (2000). 4075–4083
15. Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D., Lander, E.S.: Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science, Vol. 286. (1999) 531–537